Accurate information retrieval from large corpora : Non-extended Swahili monosyllabic verbs

Hurskainen, Arvi

SALAMA - Swahili Language Manager
2021-02-05

Hurskainen , A 2021 ' Accurate information retrieval from large corpora : Non-extended Swahili monosyllabic verbs ' Technical Reports on Language Technology , no. 67 , SALAMA - Swahili Language Manager , Helsinki . < http://www.njas.helsinki.fi/salama/accurate-information-retrieval-from-large-corpora-1.pdf >

http://hdl.handle.net/10138/330101

Technical Reports on Language Technology
Report No 67, 20021
http://www.njas.helsinki.fi/salama

Accurate information retrieval from large corpora: Non-extended Swahili monosyllabic verbs[1]

Arvi Hurskainen
Department of World Cultures, Box 59
FIN-00014 University of Helsinki, Finland
arvi.hurskainen@helsinki.fi

Abstract

Accurate information retrieval in strongly inflecting languages is problematic, because the inflected forms of words make the formulation of the search key difficult. Often the result has weaknesses in recall or precision, or in both. There is strong motivation for such a search system, where we have access also to the base form of the word, and the search could be made on the basis of the base form instead of the surface form. This approach would require the analysis of the source text and further modifications for making the text suitable for search.

Small and medium-size texts could be handled in this way, but it is not feasible to use this method for large text corpora. The analysis and disambiguation routines would take too much time for making the method suitable for large texts.

There is, however, a solution to the problem. The search can be made in two phases, so that first we use the normal string search and get the sentences that contain the desired words, plus many such sentences that we do not want. In the second phase we process the sentences into such a format, where each word has also its base form included. Now the search is made using the base form as key. The result will have only those sentences that we want to have.

I test the method with Swahili mono-syllabic verbs, because they are very difficult to handle in search tasks. Swahili has a small set of monosyllabic verbs, which seem to belong to the very basic core of the language. Because the verbs may take prefixes and suffixes, it is not always easy to judge whether the word is a verb or not. Some monosyllabic verbs also have more than one meaning, which poses a challenge to disambiguation in translating Swahili to another language. The mono-syllabic verbs have also extended forms, some of which are not any more monosyllabic. They behave in a different way than the mono-syllabic versions, which also causes a problem in describing the verbs morphologically.

In this report I restrict to testing how the non-extended monosyllabic verbs can be retrieved from large text corpora. I exclude such forms, which have the final character *e* (subjunctive) or *i* (present tense negative). I also exclude extended forms.

Key Words: *information retrieval, language analysis.*

---

1 Introduction

Among the core verbs in Swahili are mono-syllabic verbs. They belong to the very old layer of words in language. They also occur in text frequently, which is why they should be handled faultlessly in language technology. The fact that the stable element in verb is just a single consonant, or in some cases a cluster of consonants, makes the identification of mono-syllabic verbs difficult.

Because mono-syllabic verbs in text have very little stable material, that is, only the stem consonant or consonant cluster, direct string search is not ideal method because of excessive noise. The result may be either over- or under-productive, but only seldom accurate.

In this report I describe a method of retrieving monosyllabic verbs accurately, although the source text is not tagged. In tests I use the non-tagged corpus of 25 million Swahili words.

The search is made in two phases. First, the normal string search is made by using such search keys, which cover all forms of the verb. Such keys will find all true hits but also many such cases, which are not monosyllabic verbs.

If the context of the hit is sentence-long, the result will consist of true sentences, which are suitable for disambiguation.

After the first round we have a set of sentences, which contains all true hits but also many unwanted sentences.

In the second phase we analyse and disambiguate the sentences and then we retrieve the verbs using the base form of the verb as search key. It is also possible to yse the root form of the verb as search key, because the analysis contains the base form and root form of each verb.

If the source text is very big, as the Swahili corpus is, it is likely that the text after the first search round is still too big to be analysed in fly. It may be advisable to remove such wrong hits, which are easily removable without risking the true hits. I will show the process below.

2 Search in two phases

Accurate information retrieval from large text corpora needs a two-phase process. First we retrieve all those sentences, which contain the targeted words. The result should have all the targeted words. In practice, however, the result includes also such sentences, which match with the search string, but the matched word does not belong to the category of intended matches. In other words, the result has all the sentences that we want, but also such sentences that we do not want.

The text that we will now deal with is much smaller than the original corpus, often just a small fraction of it. The size of the resulting text depends on what we are searching for. For example, if we search all occurrences of the verb *ku-wa* (to be), it is likely that the resulting text is still too large to be handled in the two-phase process. Such search tasks are, however, exceptional, and in most search tasks the method that we are discussing here is suitable.

Below I will test the proposed search method with some monosyllabic verbs.

2.1 The verb *ku-nya* (to fall, to drop, to shit)

With the search key `'nya '` we get 75802 hits from the Swahili corpus, far too many for using the advanced search method.[2] The reason is that in addition to the true hits we get many wrong hits, such as *ku-fanya* (to do).

We can solve the problem by removing from the first search result such common hits that are definitely wrong.

In (1) we see the first 30 hits with the search key `'nya '`.

(1)

```
$ cat-allmat | kwic 'nya ' | head -30
naye kama David Onyango, amefanya tukio la kushangaza ambalo
a lilitokea kwenye Ikulu ya Kenya Desemba 22 mwaka huu kweny
wakiwa wanamwangalia bila kufanya lolote.
. Onyango ambaye baada ya kufanya tukio hilo alishikiliwa na
cho cha hatari ambacho kilimfanya akeshe usiku huo ndani ya
                         "Alifanya maamuzi yote makubwa peke
imeripotiwa kukamatwa kwa Wakenya wawili, kuhusiana na wizi
sa hizo za CRDB, ikidaiwa Wakenya hao walikutwa na kiasi kik
                           Wakenya hao ambao bado wanashikili
bao bado wanashikiliwa huko Kenya ni Charles Njoka, 36, na K
 kipindupindu wakisema amewafanya washerehekee New Year na n
         Licha ya kuzuiwa kufanya biashara za vyakula na vin
ta iliyopita, na kwamba kinafanya pia kazi ya kupiga vita ma
       Akasema wamekuwa wakifanya kazi hiyo katika maeneo mb
             "Kwa kweli tunafanya kazi hii katika mazingira
nye vituo mbalimbali vinavyofanya kazi za Ukimwi.
kiwa na mfadhili tutaweza kufanya vizuri kazi yetu hii ya ku
gamoto ambayo itawasaidia kufanya kazi zao kwa kujiamini zai
isema kuwa TOC inatarajia kufanya mipango ili kukutana na vi
uwa mgogoro huo huenda ukawafanya viongozi wa FAT kushindwa
ma kwamba wanataka kwanza kufanya uchunguzi wa kina juu ya k
iliana na Polisi ili waje kufanya uchunguzi wao, lakini hadi
kwa Polisi hao kufika leo kufanya uchunguzi.
                Akasema watafanya hivyo kama vile walivyokub
fanyabiashara kutoka nchini Kenya ambao hununua machungwa ya
kwamba dalali yeyote atakayefanya kinyume na amri hiyo atach
oyo wa kuungana kama walivyofanya huko Kenya.
                             Kenya imemaliza uchaguzi wake hi
enya, ni kwamba wananchi wa Kenya waliwalazimisha wapinzani
is Mkapa akaombwa akubali kufanya mabadiliko ya Katiba, Tume
```

The true hits are so rare that this sample does not contain a single one. When we remove obvious wrong hits using the script in (2), we will get a much reduced text, with only 141 sentences

[2] The search key `'nya '` excludes such forms as *nyi* (present tense negative) and *nye* (subjunctive). However, these forms are not common.

(2)

```
$ cat-allmat | egrep 'nya ' | egrep -v
"(fa|Fa|Ke|ke|o|sa|o|ya|e|pa|wa|ka|ga|Sa|a|ha|ch|hu|ia|mi|ng.u|bi|Nku)nya
" | wc
```

Part of the modified result is in (3). We see that now there are many true hits but also wrong hits. The result was reformulated using `kwic`, because `egrep` does not do aligning.

(3)

```
         Mshitakiwa alibaki kinya huku amejiinamia kama anay
iti, huku Rajabu Kigundula akinya kuwa nafasi ya Ukatibu wa
yumba ya kulala wageni ya Turunya iliyoko Mwenge Jijini, Wil
        INTERVIEWEE(S): 1. Mgunya Mwinyi (MG), 2. Maalim Haj
wa sehemu husema hu... si Wagunya na kadhalika, wapo.
ima, ee. Haya, ikawa vua inakunya kunya ngia ga uko kati.
e. Haya, ikawa vua inakunya kunya ngia ga uko kati.
a MV M Bali aaah kwa vua inakunya uzuri.
                 Ka MV M Inakunya vyema.
    MB M Wazee wa zamani, matunya hamna, watu wengi sasa haw
dua kwamba, nafikiri mnao Wagunya kiasi.
 kayumila maneno pannembele kunya baaya linamba libali kulik
pane yungula kwaa fisi nkwa kunya mavi nkutukuta mpaka kulyu
                X1 F Akenda kunya haja chumbani kule akaja k
ya haja chumbani kule akaja kunya ukumbini akenda kunya bara
akaja kunya ukumbini akenda kunya barazani.
                X1 F Akenda kunya haja. Chumbani kule. Akaja
 haja. Chumbani kule. Akaja kunya ukumbini, akenda kunya bar
kaja kunya ukumbini, akenda kunya barazani, jamaa wale. Na y
isi kuzaliwa kwetu kwa mamungunya (kicheko),
oni kwa giza. Hapo ilianzia kunya mvua asubuhi, mvua iliilie
siku nyingine. Anye, akesha kunya zile zile kata-- kanapakus
ajili ya tamaa yake ya kuibiginya nchi hiyo hadi kufa.
stan na India ambavyo vimejikunya katika mpaka wao wa Kashmi
ombi ambalo limekutana na ukimnya mzito kutoka mataifa tajir
ya nchi hiyo kulaumiwa vikali nya Jumuiya ya Ulaya kwa kukan
ubwa wa mikate yanapokea vinyunya vya mikate iliyogandikwa k
na Bemba alisema waasi wa chamnya chake cha MLC walikuwa wan
```

Now the text has such size that we can analyse and disambiguate it and process further, so that it is suitable for accurate search.

If we only need to find the hits without context, we can search from the disambiguated text, because it contains base forms of words (4).

(4)

```
 $ cat-allmat | egrep 'nya ' | egrep -v
"(fa|Fa|Ke|ke|o|sa|o|ya|e|pa|wa|ka|ga|Sa|a|ha|ch|hu|ia|mi|ng.u|bi|Nku)nya
" | atta | swasent | wordlist | dis | perl -pe 's/\>"( )?\n/>" /gm' |
egrep '"<' | perl -pe 's/\t//gm' | egrep '"nya"'
"<akinya>" "nya" V 1-SG3-SP VFIN NO-SP-GLOSS COND-IF PR:na { when } z
[nya] { shit } SV @FMAINVtr-OBJ>
"<inakunya>" "nya" V 9-SG-SP VFIN { it } PR:na 1-SG2-OBJ OBJ NO-OBJ-GLOSS
z [nya] { shit } SV @FMAINVtr+OBJ>
"<kunya>" "nya" N 15-SG z [nya] { shit } SV @OBJ
```

```
"<inakunya>" "nya" V 9-SG-SP VFIN { it } PR:na 1-SG2-OBJ OBJ NO-OBJ-GLOSS
z [nya] { shit } SV @FMAINVtr+OBJ>
"<*inakunya>" "nya"
"<kunya>" "nya" N 15-SG z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" N 15-SG z [nya] { shit } SV @PCOMPL-S
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" V INF { to } z [nya] { rain } SV @-FMAINV-n
"<*anye>" "nya" V SBJN 1-SG3-SP VFIN { she } z [nya] { *shit } SV CAP
@FMAINVtr-OBJ>
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<nya>" "nya" <Heur> N 9/10-SG { nya } @OBJ
"<alinya>" "nya" V 1-SG3-SP VFIN NO-SP-GLOSS PAST z [nya] { shit } SV
@FMAINVtr+OBJ>
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<kunya>" "nya" N 15-SG z [nya] { shit } SV @<P
"<tuliokunya>" "nya" V 2-PL1-SP VFIN NO-SP-GLOSS PAST 2-PL-SUB-REL { who
} INFMARK z [nya] { drop } SVO MONOSLB @FMAINVtr-OBJ>
"<nya>" "nya" <Heur> N 9/10-SG { nya } @OBJ
"<kunya>" "nya" V INF { to } z [nya] { shit } SV @-FMAINV-n
"<*ikinya>" "nya" V 4-PL-SP VFIN NO-SP-GLOSS COND-IF PAST { when } z
[nya] { *shit } SV CAP @FMAINVtr-OBJ>
"<*kunya>" "nya" V INF { to } z [nya] { *shit } SV CAP @-FMAINV-n
"<nya>" "nya" <Heur> N 9/10-SG { nya } @OBJ
"<nya>" "nya" <Heur> N 9/10-SG { nya } @SUBJ
"<anakunya>" "nya" V 1-SG3-SP VFIN { he } PR:na 1-SG2-OBJ OBJ { you } z
[nya] { shit } SV @FMAINVtr-OBJ>
"<ikinya>" "nya" V 9-SG-SP VFIN NO-SP-GLOSS COND:ki z [nya] { shit } SV
@-FMAINV-n
"<ilikunya>" "nya" V 4-PL-SP VFIN { they } PAST INFMARK z [nya] { rain }
SV @FMAINVintr
"<nya>" "nya" <Heur> N 9/10-SG { nya } @SUBJ
"<wanaokunya>" "nya" V 2-PL3-SP VFIN NO-SP-GLOSS PR:na 2-PL-SUB-REL { who
} INFMARK z [nya] { drop } SVO MONOSLB @FMAINVtr-OBJ>
"<wanaokunya>" "nya" V 2-PL3-SP VFIN NO-SP-GLOSS PR:na 2-PL-SUB-REL { who
} INFMARK z [nya] { drop } SVO MONOSLB @FMAINVtr-OBJ>
"<anye>" "nya" V SBJN 1-SG3-SP VFIN NO-SP-GLOSS z [nya] { shit } SV
@FMAINVtr+OBJ>
"<akinya>" "nya" V 1-SG3-SP VFIN { she } COND-IF FUT:ta { when } z [nya]
{ shit } SV @FMAINVtr+OBJ>
"<anye>" "nya" V SBJN 1-SG3-SP VFIN NO-SP-GLOSS z [nya] { shit } SV
@FMAINVtr+OBJ>
"<akinya>" "nya" V 1-SG3-SP VFIN { he } COND-IF PAST { when } z [nya] {
shit } SV @FMAINVtr+OBJ>
```

If we also want context, the process is more complicated but manageable. The result of (2) above will be analysed, disambiguated, and processed, so that it has the format as in (5). In order to save space, only part of result is displayed.

(5)

```
 Kukojoa {kojoa_V} na {na_CC} kunya {nya_V} kitandani {kitanda_N} : Mtoto
{mtoto_N} anapokojoa {kojoa_V} kitandani {kitanda_N} baada_ya
```

```
{baada_ya_PREP} miaka {mwaka_N} miwili {wili_NUM} isionekane {onekana_V}
kama {kama_ADV} jambo {jambo_N} la_ajabu {la_ajabu_ADJ} .
 Maaluni {maaluni_N} hawa {hawa_PRON} hata {hata_ADV} huko {huko_ADV}
kwenye {kwenye_PREP} mwezi {mwezi_N} watakakokwenda {kwenda_V} siku
{siku_N} wakikuta {kuta_V} mavi {mavi_N} watasema {sema_V} sisi
{sisi_PRON} ndiyo {ndiyo_ADV} tuliokunya {nya_V} huko {huko_ADV} .
 Kivunya {Kivunya_Heur} alisema {sema_V} kabla {kabla_ADV} hawajamvamia
{vamia_V} , watu {mtu_N} hao {hao_PRON} walidhibitiwa {dhibitiwa_V} na
{na_CC} kupata {pata_V} kipigo {kipigo_N} kutoka_kwa {toka_kwa_PREP}
walinzi {mlinzi_N} wa {wa_GEN-CON} Bw {Bw_N} .
 Alisema {sema_V} hayo {hayo_PRON} akikumbuka {kumbuka_V} mambo {jambo_N}
ya {ya_GEN-CON} utoto {utoto_N} alipokuwa {wa_V} akirudi {rudi_V} usiku
{usiku_ADV} siku {siku_N} za {za_GEN-CON} Jumamosi {jumamosi_N} alipoanza
{anza_V} mapenzi {penzi_N} na {na_CC} Jonas {Jonas_Heur} mpaka
{mpaka_CONJ} alipomwadhibu {adhibu_V} kwa {kwa_PREP} kumfinya {finya_N}
na {na_CC} kisu {kisu_N} kwenye {kwenye_PREP} mapaja {paja_N} .
 Mkurugenzi {mkurugenzi_N} huyo {huyo_PRON} aliwataka {taka_V} wananchi
{mwananchi_N} kujihadhari {jihadhari_V} na {na_PREP} ugonjwa {ugonjwa_N}
huo {huo_PRON} na {na_CC} kuacha {acha_V} kunya {nya_V} juisi {juisi_N}
mitaani {mtaa_N} .
 Mama {mama_N} atuhumiwa {tuhumiwa_V} kumfinya {finya_V} kwa {kwa_PREP}
ncha {ncha_N} ya {ya_GEN-CON} kisu {kisu_N} ukeni {uke_N} bintiye
{binti_N} mdogo {mdogo_N} .
 Aliwataja {taja_V} watuhumiwa {mtuhumiwa_N} hao {hao_PRON} kuwa
{kuwa_CONJ} ni {ni_V} Patrick {Patrick_Heur} Mahinya {Mahinya_Heur} ( 35
{35_NUM} ) na {na_AG-PART} Mapesa {pesa_N} Mkwavi {Mkwavi_Heur} ( 17
{17_NUM} ) .
 Kunya {nya_V} ni {ni_V} aula {la_V} zaidi {zaidi_AD-ADJ} kwa {kwa_PREP}
kukatazwa {katazwa_V} mwingine {ingine_ADJ} hakihitajii {hitajia_V} .
```

We see that there are words with the stem {nya_V} but also other stems ending with the characters *nya*.
When we use an advanced search system, we can retrieve only those sentences, which contain the stem *nya* (6).

(6)

```
 Haya, ikawa vua inakunya {nya_V} kunya ngia ga uko kati.
       -----
 Bali aaah kwa vua inakunya {nya_V} uzuri.
       -----
 Akenda kunya {nya_V} haja chumbani kule akaja kunya {nya_V} ukumbini
akenda kunya {nya_V} barazani.
       -----
 Akaja kunya {nya_V} ukumbini, akenda kunya {nya_V} barazani, jamaa wale.
       -----
 Hapo ilianzia kunya {nya_V} mvua asubuhi, mvua iliiliendelea mpaka
ikafika saa sita.
       -----
 Anye {nya_V}, akesha kunya {nya_V} zile_zile kata -- kanapakusiwa,
usubuhi kazi iwapo.
       -----
 Vyama vingine vinavyofikiriwa kushirikiana na chama cha leba katika
serikali ni MERETS kilichoshinda viti 9, Yisrael-ba alinya {nya_V} chama
ambacho ngome kuu ni wahamiaji ambacho kina viti 7.
       -----
```

```
 Kukojoa na kunya {nya_V} kitandani: Mtoto anapokojoa kitandani baada_ya
miaka miwili isionekane kama jambo la_ajabu.
       -----
 Maaluni hawa hata huko kwenye mwezi watakakokwenda siku wakikuta mavi
watasema sisi ndiyo tuliokunya {nya_V} huko.
       -----
 Mkurugenzi huyo aliwataka wananchi kujihadhari na ugonjwa huo na kuacha
kunya {nya_V} juisi mitaani.
       -----
 Kunya {nya_V} ni aula zaidi kwa kukatazwa mwingine hakihitaji.
       -----
  Bila kusita niliwapa wazi kuwa wao ni wezi na mafisadi wanaokunya
{nya_V} kwenye ofisi zetu huku wakijiona wajanja.
       -----
Kunya {nya_V} anye {nya_V} kuku, akinya {nya_V} bata kaharisha.
```

We managed to retrieve the occurrences of the monosyllabic verb *nya* from the corpus of 25 million words.

2.2 The verb *ku-cha* (to rise, to fear)

We make another test with the verb *ku-cha* (to rise, to fear). This task is much more difficult, because the string *cha* is also a genitive connector in class 7 (e.g. *kitabu cha mtoto*, child's book). When our key word in search is `'cha '`, we get 123995 occurrences, that is, vastly more than we can conveniently analyse.
We take a look at the result, so that we can see, how we could safely reduce the result (7).

(7)

```
        Chuo cha Taifa cha Utalii Nchini kimepanga ku
huo kinachotumiwa kiko kimoja cha Forodhani ambacho hivi sas
hatua zinazochukuliwa na chuo cha Taifa cha Utalii katika ku
zochukuliwa na chuo cha Taifa cha Utalii katika kuingia kati
a na ukweli kuwa hakuna kikao cha Kamati ya Utendaji ya klab
ilisema kama kweli kuna kikao cha Kamati ya Utendaji iliyoku
           Mtibwa ilitaka kumuacha Costa ikiwa tu Simba itaku
                  Kipigo chake cha kwanza ni kuiteka Dar es s
                         Chama cha upinzani cha CUF kimeanza
            Chama cha upinzani cha CUF kimeanza mikakati ya n
                               Licha ya ajenda hiyo ya maombole
e umuhimu wa kuipa CCM kipigo cha aibu katika uchaguzi mkuu
jeruhiwa vibaya na wengi wakaacha makazi yao na kutimkia huk
 ya maandamano katika kipindi cha takribani wiki moja tu.
a yaliko Makao Makuu ya chama cha upinzani cha TLP, chama am
o Makuu ya chama cha upinzani cha TLP, chama ambacho kimekuw
aka mgambo waliompiga mpiga picha waadhibiwe vikali.
                         Chama cha Wapigapicha za Habari Nchi
             Chama cha Wapigapicha za Habari Nchini, PPAT, ki
rifa hiyo akasema kuwa mpigapicha huyo alikuwa katika majuku
inondoni kukifuatilia kikundi cha mgambo hao kilichofanya sh
ya shambulio dhidi ya mpiga picha huyo na kukichukulia hatua
lipokuwa katika eneo la kituo cha mabasi cha Ubungo ambapo l
tika eneo la kituo cha mabasi cha Ubungo ambapo licha ya kuw
ha mabasi cha Ubungo ambapo licha ya kuwaonyesha kitambulish
sha kitambulisho chake halali cha kazi kinachotolewa na Seri
```

```
bia za kunyanyasa wagonjwa kuacha mara moja.
ea maisha bora zaidi, kwani licha ya kupata mshahara kazini
ayo imefanya kazi kwa kipindi cha miaka mitano, imekusanya p
u, umehifadhiwa katika chumba cha kuhifadhia maiti katika ho
a kuongeza kuwa marehemu hakuacha ujumbe wowote na wala saba
toka kilabuni kupata kinywaji cha pombe za kienyeji.
```

We see that we can remove the string `' cha '` (space on both sides), because the only context, where the form could occur as a verb is imperative singular, which is very unlikely. When we do this, we get the result as in (8).

(8)

```
aaluma mkoani, hivyo kwa kuwaacha walimu hao kutoshiriki zoe
ika ajali hiyo ambapo pia imeacha yatima na pia kuwaacha wen
ia imeacha yatima na pia kuwaacha wengine wakiwa na vilema v
ngeta Kasika (54), ambaye ameacha watoto wanane ambao waliku
li hiyo ambapo Rhoda naye aliacha watoto watatu.
                 Marehemu wameacha watoto ambao walitegemea k
      Ajali hii mbaya pia iliacha Watanzania kila mmoja akiw
Jeshi Stars, iliyosaniwa na kocha wake mkuu, Martin Kemwaga
                             Licha ya Bunge kutangaza kumpa m
 wanalipwa mafao yao mapema licha ya kasoro iliyojitokeza,"
  Kwa mujibu wa Bw. Nyundo, licha ya mchango wa Manispaa, Se
                      Aibu kuikacha Simba - Asili.
ji takribani watano enzi za kocha wao Mkongo, Raoul Shungu.
walivyosema kuwa Yanga imewakacha ni sahihi kabisa.
                             Kocha wa timu hiyo, William Kima
                         Mkijificha kwenye makaburi, nani ataw
wakienda huko, lazima watajificha tu wakiogopa kukamatwa," a
 amewaasa viongozi wa dini kuacha kutoa kauli zinazoweza kuh
inyonga na katika ujumbe alioacha amedai kuwa amejimaliza ku
asema kabla ya kujinyonga ameacha ujumbe unaosema kwamba ame
                          Nimeacha mke na mtoto mmoja.
 amewataka wakazi wa Sinza kuacha dharau kwa kuwa gonjwa hil
                             Licha ya kumuua mfanyabiashara h
                         Mpigapicha Kassim Makongoro, 48, anay
na angependa Kassim ampigie picha mbili za kawaida toka kwen
                      Alitaka picha anayoingia ukumbini na ali
na siku zinakwenda na haoni picha zake zikiletwa, alinieleza
wa makubalianio ya kumpigia picha za kawaida toka kwenye mka
ni wamewataka viongozi wao kuacha kuzungumza uongo katika vy
ichezo Taifa, BMT, kutaka makocha wote akiwemo yeye kuwasili
eye kuwasilisha CV zake za ukocha IFCA.
ibuni BMT ilitangaza kuwa makocha wote Tanzania nzima wajian
a CV zao katika vyama vya makocha vya wilaya ili waweze kuwa
            Wakati huo huo, kocha Boniface Mkwasa tayari ame
la aliyasema hayo jana kuwa kocha huyo amepeleka vyeti juzi
John Pombe Magufuli leo amewaacha wabunge midomo wazi pale a
     Akasema mwanamama huyo licha ya kubembelezwa na mumewe
aga michezo hiyo, na hivyo kuacha kibarua kwa Mtagwa ambaye
wa kupata medali, na hivyo kuacha kazi kwa Mtagwa na Aziz Mw
                             Licha ya kumuua mfanyabiashara h
```

Here we can see, how we could do further pruning. We remove such strings, which cannot be forms of the verb 'cha' (9).

(9)

```
$ cat-allmat | kwic 'cha ' | egrep -v ' (cha|\*cha) ' | egrep -v '((
|,)li|Li|[Kk]o|wa|a|[Pp]i|fi|[Vv]o|hi|n|uku|aku|do|vi|bu|
m|fiki|e|ti|A)cha ' | wc
    985    8820   57745
```

In the result we have 985 such sentences, which include the verb *cha* and also some such sentences that we do not want to get. This text is small enough for processing through the second phase. When we do this, we get 604 such sentences, where the string *cha* is a monosyllabic verb. Part of result is in (10).

(10)

```
 Firauni alipotoa agizo la_kuwaua wazaliwa wa_kwanza wa_kiume, wakunga wa
kiyahudi walikataa kutii agizo lake: "wale wakunga walikuwa wakimcha
{cha_V} Mungu, wasifanye kama walivyoamriwa na huyo mfalme wa Misri,
lakini wakawahifadhi hai wale watoto wanaume" ( Kut.1:17 ).
       -----
 Anasema dunia ya_sasa ni ya utandawazi, hivyo kila kukicha {cha_V} njia
za utendaji zinabadilika, hivyo ni wajibu wa kiongozi kuendana na hali
hiyo.
       -----
 Kila kukicha {cha_V} matukio ya vifo, kujeruhi kwa silaha za_hatari kama
visu, bastola baina_ya wapenzi ( mke na mume ) au watu wenye uhusiano
kibiashara au kifamilia yamekuwa yakitokea.
       -----
 cha kuwashitua Watanzania hali iliyosababisha Serikali kuunda Tume
kuchunguza nini hasa KILA kukicha {cha_V} katika vituo vya daladala
jijini Dar_es_Salaam hukosi kukutana na ndugu zetu wanaoitwa vibaka au
mateja Heur V.
       -----
 Hii inatokana_na utaratibu mbovu na wa_kiuonevu na usiotoa haki kwa
aliyekosa au aliyekosea na hivyo kusababisha migogoro kuendelea kila
kukicha {cha_V} na kudumaza michezo.
       -----
 Je, iko wapi ripoti ya jopo la Kova? Kwenye hotuba zenu kila kukicha
{cha_V} tunasikia tu habari za mipango ya kuweza kukopesha wakulima.
       -----
 Je, iko wapi ripoti ya jopo la Kova? Kwenye hotuba zenu kila kukicha
{cha_V} tunasikia tu habari za mipango ya kuweza kukopesha wakulima.
       -----
 Kumekucha {cha_V} Ligi Kuu Tanzania_Bara 2013/14 Lyamunda anasema
wananchi wanapaswa kujiuliza ni faida gani wameipata tangu kuanza
kuchimbwa kwa madini ya almasi, tanzanite na mengineyo ambayo kila
kukicha {cha_V} yanazalisha migogoro mingi iliyokosa utatuzi mpaka
hivi_sasa.
       -----
 Serikali inapaswa kulaumiwa kwa wizi unaoendelea kutokea kila kukicha
{cha_V} kwenye wizara, idara, taasisi na mashirika yaliyo chini_yake,
kwa_sababu mara nyingi imeshindwa kuwachukulia hatua kali za_kinidhamu na
kisheria wanaobainika kutenda ufisadi.
       -----
 Anasema wamekuwa wakiwapatia misaada ya fedha, nguo na hata vyakula kila
kukicha {cha_V} hivyo huruma yao ndio inawafanya wao kuendelea kuomba.
       -----
```

```
 Kumekucha {cha_V} uchaguzi DRFA Azam 'mtoto' kwa Simba Kwa_mujibu_wa
habari hizo, kucha za vidoleni na miguuni zimeoza na zina rangi nyeusi,
miguu na mikono yake imepasuka mithili_ya mtu mwenye magaga Heur ambayo
wakati mwingine hutoa damu na kumsababishia maumivu makali.
       -----
 Nyalandu alisikitikia makundi hayo kwa kutumia muda mrefu kulumbana na
kufanyiana fitina wakati nchi jirani zikipigana vikumbo kila kukicha
{cha_V} kutafuta nafasi kuja kupata ajira hapa nchini sambamba_na
kuchukua ardhi.
       -----
 Ushauri mwingine ni kwamba wanamuziki wa hapa nyumbani wengi_wao
wameamka na kujaribu tu kuimba ikiwa ni katika sehemu ya kujitafutia
riziki nasema hivyo kwa_sababu nina uhakika wengi hawana ujuzi wa_kutosha
katika tasnia hiyo ndiyo_maana kila kunapokucha {cha_V} kazi zao zina
upungufu.
       -----
 Lakini wakati tunasema hatuna pesa kwa_ajili_ya kuinua huduma za_kijamii
hasa afya na elimu, tuna pesa za_kutosha kujenga majumba ya_kifahari
kwa_ajili_ya spika, waziri_mkuu, gavana, ukumbi wa_kifahari wa Bunge
ambapo wapo wengi wanautumia kuucha {cha_V} usingizi na matumizi mengine
mengi kama vitafunwa na posho kwa kazi ambazo tayari wahusika wanalipwa
mishahara kuzifanya Tatizo letu liko kwenye kuchagua na kupanga.
       -----
 Uharamia huo umekuwa ukishika_kasi kila kunapokucha {cha_V} huku mamlaka
zinazohusika zikiendelea kupanga na kupangua mikakati ya kukomesha mtindo
huo bila mafanikio.
       -----
 Taifa likatangaziwa kuwa hata rais wa nchi alilazimika kusitisha safari
ya nchi za_nje ili ahudhurie msiba huo wa_aibu kubwa kwa wanaomcha
{cha_V} Mwenyezi Mungu Sema kwakuwa_na safari zenyewe siku_hizi zina sura
ya matembezi zaidi.
       -----
 Taifa likatangaziwa kuwa hata rais wa nchi alilazimika kusitisha safari
ya nchi za_nje ili ahudhurie msiba huo wa_aibu kubwa kwa wanaomcha
{cha_V} Mwenyezi Mungu Sema kwakuwa_na safari zenyewe siku_hizi zina sura
ya matembezi zaidi.
       -----
```

The majority of the occurrences of the verb *cha* are in the context *kila kukicha* (every time when sun is rising, that is, every morning). The extract above contains also other uses.

In the corpus there occurs the noun *ukucha/makucha* (class 11/6) in the form *kucha* (class 9/10), and this can be easily mixed with the infinitive form *kucha* of the verb. Disambiguation, however, solves the problem.

2.3 The verb *ku-la* (to eat)

There are still more monosyllabic verbs. Next we test the verb *kula* (to eat), which is common in texts. The verb stem *la* is identical with the genitive connector of class 5 (e.g. *swala la serikali* (problem of the government)). It is also the negative interjection.

Without any constraints, the search key `'la '` produces 354880 hits. This means that heavy constraining is needed.

When we use the constraints described in (11), we get a much reduced text.

(11)

```
$ cat-allmat | kwic 'la ' | egrep -v ' (la|\*la) ' | egrep -v
'([Kk]i|[Bb]i|[Mm]i|f|
i|b|da|ye|ua|gu|la|um|ha|Mo|ga|l|ge|pa|taa|ro|le|ma|me|je|swa|bu|[Dd]o|to
|[Tt]awa|[Nn]aka|Kw.|e|[Cc]haku|[Vv]yaku|sa|hi| [Ii]|lu|[Mm]awaka| m|
[Ww]a|go|[Llo]|ba|[Ll][ou]|ai|za|t|fu|su|[Mm]aka|[Mm]u?hu|
[Ww]aka|ra|di|ti|ou| [Ss]wa|[Mm]ihu|tu)la ' | wc
   2906 lines
```

These 2906 sentences will be analysed and processed into rich text format. Then search is targeted into this text using the advanced search system. This is demonstrated in (12).

(12)

```
$ echo {la_V | find-word
 Pia ikadaiwa kuwa alimwambia kwamba Bilal anakula {la_V} haja_kubwa kwa
kutumia kijiko.
       -----
 Katika sherehe hiyo, wanachama hao walikula {la_V} na kunywa na pia
kuburudika kwa muziki kuanzia saa 5:oo asubuhi hadi saa 12:oo jioni.
       -----
 Aidha diwani huyo akawataka wafanyabiashara ambao wanapeana michezo
kuacha tabia hiyo kwa_kuwa wengi wanaocheza michezo hiyo wanakuwa sio
waaminifu na wanakula {la_V} pesa za wenzao.
       -----
 Yanga kumsaka aliyekula {la_V} dola.
       -----
 Hata_hivyo, Kamanda amesema Polisi watakula {la_V} sahani moja na
mwanamke aliyemuacha_solemba mwanae.
       -----
 Abdallah Matata amesema katika kijiji chao wamekuwa wakidai mkutano wa
kijiji kwa zaidi_ya mwaka mmoja sasa lakini wapi kwani mwenyekiti anazidi
kula {la_V} henga.
       -----
 Akaongeza kuwa baada_ya kufanya mkutano na kujichagua, walisambaza
vikaratasi vya kuonyesha kuwa viongozi wa_zamani hawafai kwani wamekula
{la_V} pesa za kijiji.
       -----
 Hakimu_Mkuu Mkazi wa mahakama ya Hakimu Mkazi Kisutu, Mheshimiwa Projest
Rugazia ameula {la_V} kufuatia Rais Mkapa kumtangaza kuwa mmoja wa majaji
sita wapya wa Mahakama Kuu nchini.
       -----
 Inadaiwa kuwa hupendelea kula {la_V} mikate, ugali, wali pamoja_na
kitoweo safi.
       -----
 Wale wanaume wenye tabia ya kuwapiga wake zao sasa wakae_chonjo kufuatia
Mkuu wa Mkoa wa Dar_es_Salaam Luteni Yusufu Makamba kuahidi kuwa atakula
{la_V} sahani moja na watu wenye tabia hiyo.
       -----
 Akasema alikula {la_V} kibano hadi alipoamua kuwatajia mahali anakoweka
pesa zake, na majambazi hayo yakafanikiwa kuzoa Shilingi 600,ooo.
       -----
 Pamoja_na hayo Meya ameahidi kula {la_V} sahani moja na wazazi
wanaowaoza watoto wao wakiwa bado wanafunzi.
       -----
```

```
 Hilo likamchanganya mwanamama huyo mwenye kampuni na kulazimika kula
{la_V} ' na maofisa wa PCB ili wamsaidie kukomesha utapeli huo.
       -----
 Serikali inakula {la_V} matapishi yake Wananchi kadhaa jijini Dar
wamesema safari wako tayari kuandamana kutokana_na serikali kula {la_V}
matapishi yake kwa kupandisha umeme kupita uwezo wa kipato cha Mtanzania.
       -----
 Huku ni kula {la_V} mapishi yake yenyewe" akashangaa mama Anna.
       -----
 Kwa_nini serikali imeamua kula {la_V} matapishi yake?" akahoji Mzee
Kilindo kwa hasira.
       -----
 Kwani wanaokula {la_V} mikate si wakubwa bwana Katika kesi ya mganga
wa_kienyeji, inadaiwa na mwendesha_mashitaka, mrakibu mwandamizi wa
polisi, ASP Willy Mlulu kuwa Februari 13 mwaka huu, hapa Jijini,
washitakiwa hao walikula {la_V} njama za kutenda wizi.
       -----
 Uoteshaji nyasi kingoni mwa barabara kula {la_V} mil "mil".
       -----
 Kamanda Tibaigana akasema mwanamama huyo, ambaye kwa_sasa ni mtuhumiwa,
alichukua uamuzi huo wa kujikatakata na kisha kula {la_V} sumu kwa_siri
kubwa.
       -----
 Katolila aliwataja wachezaji wanatakiwa kuripoti kuwa ni Hafidhi,
Mustapha Mkumba, Kijangwa Kondo, Mohamed Liusa, Said Nachikongo, Jamuhuri
na Samson Kilalo {la_V}.
       -----
 Mukama hakutaja ununuzi huo utakula {la_V} kiasi_gani { what extent }
@ADVL cha pesa.
       -----
 Christopher Mgaya, amesema wao kama vijana wasiokuwa_na ajira ambao
awali walikuwa kula {la_V} kulala, wamefikia uamuzi huo ili wapate ajira
ya_muda.
       -----
 Sisi ni wasanii, tumepitia sehemu nyingi, tumekumbana na matatizo kibao,
sasa hapa tumeamua kutulia, tunakula {la_V} raha, kwa_nini tusumbuke
sumbuke," alisema.
       -----
 Hata_hivyo, Alasiri lilipotaka kujua kwanini tajiri huyo anauza unga huo
hadharani, likaelezwa kuwa zungu huyo anakula {la_V} na mapolisi.
       -----
 Kwa_mujibu_wa taarifa za madaktari, ugonjwa huo hutokana_na kula {la_V}
nyama ya ng'ombe mwenye kichaa.
       -----
 Baada_ya kusikia hayo, ndipo DC huyo alipoamua kuwapasha wanakijiji hao
kwa kusema kuwa hata wakienda porini kwa siku sitini, bado atakula {la_V}
nao sahani moja hadi hapo mashamba yote yanakuwa meupe.
       -----
 Habari hizo pia zimethibitishwa leo na Kamanda_wa_Polisi Mkoa wa
Dar_es_Salaam, Alfred Tibaigana, ambaye kwa_uwazi na masikitiko akasema
hivi : Kwamba mkazi huyo wa Jiji alikula {la_V} samaki huyo Julai 12,
Mwaka huu saa 1:15 huko Kunduchi Pwani wilaya ya Kinondoni, Jijini.
       -----
 Mwanajeshi huyo kwa kutumia mbinu za medani alikula {la_V} sahani moja
na mjaluo ambaye baada_ya kuona anakamatwa alitoa kisu.
       -----
```

```
 Wakiendelea wawaambia makarani hao kuwa watakula {la_V} kiapo
kwa_ajili_ya kutunza siri za watu watakaowahesabu katika zoezi hilo na
atakayevunja atashitakiwa.
       -----
 Wakiendelea wamewaambia makarani hao kuwa watakula {la_V} kiapo
kwa_ajili_ya kutunza siri za watu watakaowahesabu katika zoezi hilo na
atakayevunja atashitakiwa ikiwa_ni_pamoja_na kunyimwa posho.
       -----
 Akasema majeruhi wapo wamelazwa katika wodi ya Kibasila
Hospitali_ya_Taifa_ya_Muhimbili na hali {la_V} zao zinaendelea vizuri.
       -----
 Kweli mvumilivu hula {la_V} mbivu.
       -----
 Katibu huyo akasema kuwa chama kimechoka kusikia juu_ya tuhuma
zinazowakabili walimu kuwa wamekula {la_V} fedha za sensa na mambo
mengine mengi.
       -----
 Likimangira na wenzake, upande wa mashitaka umeiambia mahakama kuwa
anadaiwa kula {la_V} njama za kutenda udanganyifu huo kuanzia Machi mwaka
juzi hadi Septemba mwaka huu muda usiojulikana.
       -----
 Akasema 'hilo kwa_sasa linakula {la_V} nyumba za watu ambazo zote zipo
katika viwanja halali vilivyopimwa na serikali.
       -----
 Mkurugenzi wa bendi hiyo, Felician Chaula {la_V} alisema jana kuwa
maandalizi ya uzinduzi huo yamekamilika hasa baada_ya vijana wake
kumaliza kambi iliyowekwa kwenye hoteli ya Bahari Beach.
       -----
 Mbunge huyo akawataka watoto kula {la_V} kitabu sawasawa ili waweze
kujiandalia maisha mazuri baadaye na hivyo kukipa faraja kiwanda hicho.
       -----
 Eti kamla {la_V} roba FFU na kisha kumpora.
       -----
 Hivi_karibuni wajumbe watano wa chama_cha_upinzani cha TLP waliokuwa
wamepangwa kutembelea ofisi ya mbunge wa huyo walidai kuwa mheshimiwa
huyo amewala {la_V} chenga ya mwili.
       -----
 Pepe kale wa CCM asema mwereka aliokula {la_V} umemfundisha uvumilivu.
       -----
 Majid akasema kinyume_chake kuangushwa kwake kumemfanya awe_na uvumilivu
huku akitegemea kula {la_V} mbivu hapo baadaye.
       -----
 Katibu wa timu ya soka ya Simba, Kasimu Dewji amesema kuwa atakula
{la_V} sahani moja na Chama cha Soka Tanzania (FAT), endapo chama hicho
kitamfungia mchezaji Mkenya Mark Sirengo.
       -----
```

2.4 The verb *ku-fa* (to die)

The search key `'fa '` yields 77374 hits, which is less than with most other monosyllabic verbs. Yet it must be reduced to make advanced search convenient. Part of the result is in (13).

(13)

```
                          Taarifa hiyo licha ya kuwatakia Wai
ki Meya wa Mansipaa hiyo Mustafa Yakub, Mkurugennzi wa Mansi
le na kuiagiza Manispaa itoa ofa ya kiwanja kingine kwa ajil
iaji wa zamani wa Yanga na Taifa Stars, Abeid Mziba amesema
tendo cha kuhujumu timu ya taifa ni cha kulifedhehesha taifa
aifa ni cha kulifedhehesha taifa na soka kwa ujumla.
                          Taarifa hizo zimepatikana baada ya
ini kiongozi wao hakutoa taarifa baada ya kuona mwenzake har
jenzi Bw. John Magufuli kwa sifa kemu kemu.
umbe wa Halmashauri Kuu ya Taifa ya CCM, NEC, ambapo mara mb
lisema kama lingetekelezwa maafa ya kivuko hicho yasingetoke
kwa wananchi waliogiswa na maafa hayo wakati yeye akiwa nje
fanikio ni kitu gani mpaka maafa yanatokea.
makao makuu wa kitengo cha maafa Dar es Salaam kuchukua sehe
cha maiti cha Hospitali ya Taifa ya Muhimbili.
 jana walifanya mtihani wa Taifa na kufaulu wote ambapo wali
         Mkazi mwingine Bi Arafa Ramadhani akamuunga mwenzie
ika kuivaa Yanga Uwanja wa Taifa bila ya nyota wake wawili,
                          Taarifa za urithi zatakiwa kuhifadh
anya juhudi za kukusanya taarifa mbalimbali za urithi wa uta
haja kubwa ya kuziingiza taarifa za aina hii, katika kumbuku
   Mbali na athari hiyo, taarifa hiyo ikasema picha zinazoon
, amesema kuwa ili timu ya Taifa iweze kufanya vizuri lazima
hagua wachezaji wa timu ya taifa liachwe kwa kocha.
izuri ama vibaya lawama ama sifa zibaki kwa kocha mwenyewe t
upata wachezaji wa timu ya Taifa siyo mzuri ambapo kamwe FAT
buni FAT iliivunja timu ya Taifa baada ya kupata kipigo cha
abao 5-0 kutoka kwatimu ya Taifa ya Kenya Harambee Stars.
ia Simba kutumia uwanja wa Taifa ni sawa na kuingiza siasa m
asaga akasema ni kukagua taarifa za vikao vya chama katika n
```

We remove obvious wrong hits (14).

(14)

```
$ cat-allmat | egrep 'fa ' | egrep -v
'(ai|ri|l|di|ra|aa|sa|hi|[Ss]i|ru|sh|o| [Uu]|ta|li|yu|fa|d|dh|ti)fa ' |
wc
6674 lines
```

We convert this reduced text into rich text format, and then we use the advanced search method. Part of result is in (15).

(15)

```
$ echo {fa_V | find-word

Aliwataja waliokufa {fa_V} katika ajali hiyo kuwa ni Fadhil Bakari (27)
wa Mbwewe mkoani_Pwani na Abdallah Ramadhani (30) wa Usangi mkoa wa
Kilimanjaro, ambao alisema miili yao ilihifadhiwa katika hospitali ya
KCMC, Moshi ikingoja kusafirishwa.
       -----
 Kamanda Ntobi alisema tukio hilo lilitokea Novemba 9 mwaka huu saa 7
usiku katika kijiji hicho wilayani Tarime ambapo alipigwa na kitu kizito
usoni na kufa {fa_V} papo_hapo.
       -----
```

Wakati_huo_huo, watu wawili wamekufa {fa_V} na wengine watano kujeruhiwa baada_ya kuvamiwa na majambazi.

Waliokufa {fa_V} katika tukio hilo ni Calvin Chiganga (21) ambaye ni mwanafunzi wa kidato cha tatu katika shule_ya_sekondari Mwembeni mjini hapa, na Song'ora Muga (55) aliyeuawa wakati akienda kutoa msaada kwa waliovamiwa.

Kamanda Kabogota alisema mtuhumiwa alimpiga rungu mwanaume katika paji la uso na kufa {fa_V} papo_hapo.

Katika tukio lingine, Bibi Cecilia Matembo (42) ameokotwa ndani_ya Mto Lumeme akiwa amekufa {fa_V} huku mwili wake ukiwa_na majeraha ya kukatwa na kitu kikali.

Cosmas Nyani (50), amekutwa chooni akiwa amekufa {fa_V} huku akitoka damu puani.

Edgar Mkane (9), amegongwa na gari na kufa {fa_V} baada_ya kuruka kutoka_kwenye baiskeli aliyobebwa.

Kamanda Kombe alisema dereva Calist Gabriel (42) alipigwa_risasi kwapani na kufa {fa_V} papo_hapo na watu wasiojulikana.

Mtoto afa {fa_V} kwa moto.

Aliongeza kuwa katika tukio jingine kwenye kambi ya Lukole Ngara, Novemba 14 saa tatu usiku, mtoto mwenye umri wa mwaka mmoja na_nusu aliyetambuliwa kwa_jina la Muhimudu Anastiata, alikufa {fa_V} kwa kuungua moto.

Watu 4 wafa {fa_V} katika ajali tofauti.

WATU wanne wamekufa {fa_V} katika matukio ya ajali yaliyotokea jana na juzi, Dar_es_Salaam.

Tukio la_pili lililotokea juzi saa 5.20 usiku katika barabara ya Temeke ambapo mwanamume anayekadiriwa kuwa_na umri wa kati_ya miaka 27 na 30 aligongwa na gari na kufa {fa_V} akitoka Chang'ombe kuelekea barabara ya Kilwa.

Katika tukio la_tatu, mwanamume mwingine anayekisiwa kuwa_na umri huohuo ambaye pia hajatambuliwa aligongwa na gari lisilofahamika jina wala namba na kufa {fa_V} jana saa 7 usiku katika barabara ya Morogoro eneo la Manzese.

Mchimba madini afa {fa_V} kwa kufunikwa na kifusi.

Kessy Mbaya, ambaye umri wake haukutajwa ameangukiwa na kifusi na kufa {fa_V} wakati akichimba madini.

Kamanda Msika alisema baada_ya wazazi wa mtoto huyo kumtafuta, walimkuta porini amekufa {fa_V}.

Habari kutoka_kwa Kamanda_wa_Polisi wa Mkoa wa Tanga, Bibi Faith Amour, zilisema watoto wawili ndugu wamekufa {fa_V} kwa kuungua moto.

```
 Aliwataka vijana kuachana na ngono, kwani vijana wengi hivi_sasa
hawafikishi umri wa miaka 40 na hufa {fa_V} mapema kutokana_na kuugua
ugonjwa wa UKIMWI.
       -----
 Watatu wafa {fa_V} kwenye matukio tofauti.
       -----
 Watu watatu wamekufa {fa_V} na wengine 32 kujeruhiwa katika ajari tatu
tofauti zilizotokea katika Mkoa_wa_Pwani wiki hii.
       -----
 Kundi hilo la majambazi lilivamia nyumbani kwa mashaka Paulo na
kufanikiwa kumpora tshs 15,000/= na baadae kurejea tena kwa Kanyarwanda
yakampiga Risasi ya kifuani na kufa {fa_V} papo_hapo.
       -----
 Watu 6 wafa {fa_V} Krismasi Dar_es_Salaam.
       -----
 WATU sita wamekufa {fa_V} Dar_es_Salaam katika matukio tofauti wakati wa
Siku Kuu ya Krismasi.
       -----
 Tukio la_kwanza lilitokea juzi saa 8.30 mchana, katika barabara ya
Uhuru, Buguruni Malapa, baada_ya gari namba TZQ 7208 Toyota Hiace,
lililokuwa likiendeshwa na Zacharia Abdu (21) kumgonga mwenda kwa_miguu
Hamis Saidi (28) na kufa {fa_V} papo_hapo.
```

2.5 The verb *ku-pa* (to give)

The verb *kupa* is a ditransitive verb. It usually occurs with the indirect object and direct object, in this order. Because the indirect object is usually an animate, it is marked also in the verb, although the sentence may contain also an overt indirect object. An example is in (16).

(16)
Nilimpa kitabu. I gave him/her a book.
Nilimpa mtoto kitabu. I gave the child a book.

The search key `'pa '` yields 53934 hits from the Swahili corpus. Extract from the result is in (17).

(17)

```
 kutozungumza na mtu yeyote hapa duniani isipokuwa mtu mmoja
            Pili, ni mwiko kumpa mkono wake mtu yeyote isipo
i sasa mambo yake ni safi akampa disichaji na kumruhusu aend
Ngulumba anayeishi Jangwani hapa Jijini.
 kutoka katika makampuni 23 hapa nchini.
fanyika Agosti 10 mwaka huu hapa Jijini kwenye Jumba la Utam
babisha kuvunjika kwa amani hapa nchini.
wa kingeweza kuvuruga amani hapa nchini.
a ya Jamhuri ya Muungano kinampa nguvu ya kulalamika anapoon
hochea kuvurugika kwa amani hapa nchini lakini tangu mwaka 1
                        Machupa asakwa na polisi.
zamani wa Simba, Athuman Machupa pamoja na watu wengine wawi
kia Alasiri zinadai kuwa Machupa akiwa na mtu aliyetajwa kam
polisi cha Mwananyamala Kwa Kopa ambako walipewa Hati Na. KP
                 Walipofika hapa na kujieleza, tuliwapa hati
```

```
fika hapa na kujieleza, tuliwapa hati nyingine Na. OB/RB/918
     Salum alisema Heri, Machupa na Hatibu waliongoza kundi
ukumbi wa Chuo cha Ushirika hapa Moshi alipomwaga kilio kika
    Amesema hivi sasa Rais Mkapa ni rafiki yake na anakipend
enge ambazo zitapita mkoani hapa na mwenge huo kulala Mzumbe
a zinachongwa, uje na upite hapa darajani ili hili daraja am
watu, kazi ambayo ilikuwa ikimpa riziki yake.
aye hakuwa na makazi maalum hapa Jijini, alikuwa habanduki n
oyote yale pamoja na wao kuogopa kuhoji hali hiyo kwani wana
ile kilichoelezwa kuwa waliogopa kukamatwa na polisi kufuati
hayo wamekimbilia porini kuogopa wasikamatwe na vyombo vya d
```

There are a few true hits, but most of them are not wanted. We constrain the search key and convert the text into sentence-per-line format (18).

(18)
```
$ cat-allmat | egrep 'pa ' | egrep -v '(go|[Hh]a|ka|we|li|hu| |[Mm]aki|
[Kk]i|ko|a|hi|fu|kutu|watu|itu|zitu|ti)pa ' | atta | swasent | egrep 'pa
' > testtext
```

The text has now 6568 sentences, a little bit too much but manageable. We covert this text into rich text format and make an advanced search (19).

(19)
```
$ echo {pa_V | find-word

 Fatuma naye akakubaliana na ombi la Mzee Abdallah na wazazi wake wakampa
{pa_V} baraka zote baada_ya bwana_harusi huyo kukata kitita cha Shilingi
150,000 kama mahari ya bibie.
       -----
 Ikadaiwa kuwa Mwajab wa makusudi aliingia_mitini na shilingi 70,000 za
Suleiman Rashid ambazo alimpa {pa_V} ili akanunuli Heur vinywaji katika
grosari yake.
       -----
  Mechi hiyo iliandaliwa na chama cha soka wilaya Ilala, IDFA, ili kuzipa
{pa_V} mazoezi timu hizo kabla hazijaenda Uganda kushiriki michuano ya
Klabu Bingwa Afrika_Mashariki na Kati, ambayo Simba ni bingwa mtetezi.
       -----
 Magoba akasema hakuona sababu ya kutompa {pa_V} kura yake kwani hata
asingempigia isingesaidia kitu.
       -----
 Mtu huyo anayedaiwa kufanya usaliti huo ametajwa kwa_jina la Hartwi
Komba, 25, mkazi wa Ubungo Kibo ambaye walikuwa wamempa {pa_V} kazi ya
udereva hapo kanisani.
       -----
 Mswaki aipa {pa_V} ushindi Faru Dume.
       -----
 TTCL yaipa {pa_V} shule Kimara madawati 50.
       -----
 Wakati kule Temeke Bwana Seif Salum Mkamba ameibuka kuwa Naibu Meya wa
Manispaa ya Temeke, Kinondoni wamempa {pa_V} nafasi hiyo Bwana Abubakari
Miringo wakati Ilala aliyeibuka_kidedea ni diwani wa kata ya Mchikichini
Bwana Mussa Azzan ama '.
       -----
```

```
 Akizungumza katika kambi ya mazoezi ya timu ya Dar_es_Salaam, Kocha Mkuu
wa timu ', William Lyimo alisema jana kuwa Nasser awali hakuwemo katika
timu hiyo, lakini sasa ameongezewa ili kuipa {pa_V} nguvu zaidi timu yao.
       -----
 Katika mchezo mwingine wa uzito huo, mwamuzi Juma Seleman alimpa {pa_V}
ushindi bondia Mayanja Brian wa Kampala katika raundi ya_nne, baada_ya
Okomu Omondi wa Mombasa kuonywa mara kadhaa na kurudia ndipo mwamuzi huyo
akalivunja pambano hilo.
       -----
 Kijana huyo amewaomba watu binafsi, mashirika, taasisi mbalimbali
kujitokeza kumpa {pa_V} msaada wa hali na mali ili aende huko India
akapatiwe matibabu.
       -----
 Taarifa hizo pia zimekuwa mpya kwa Mkuu huyo wa Wilaya, kwa_kuwa
baada_ya kuzisikia amemtaka diwani huyo kumpa {pa_V} muda zaidi
kufuatilia jambo hilo kwa_kuwa hadi_sasa hajafahamishwa ujenzi huo wa
mitaro.
       -----
 Tau akasema kuunda muungano huo kutakipa {pa_V} chama cha CCM muda wa
kuuchunguza uimara huo na kisha kuuvuruga.
       -----
 Akasema bwana huyo alipojaribu kuleta za kuleta, ndipo bibie akawa mbogo
na kuanza kumpa {pa_V} kibano.
```

Each occurrence of the verb *pa* has an object prefix that refers to the indirect object.

2.6 The verb *ku-ja* (to come)

The last verb that we test with is the intransitive verb *ku-ja* (to come). It is expected that the verb is very common and therefore might not be suitable for the kind of search that we are using here. The direct search with the key `'ja '` yields 168870 occurrences, in fact, quite a big number. An extract of the result is in (20).

(20)

```
mchezo uliofanyika kwenye uwanja wa Mandela.
a kocha Chamangwana amekuwa moja ya sababu kubwa ya timu kuf
azuri ya ushindi ikiwa ni pamoja na kuipeleka timu Malawi kw
             Amesema kuwa pamoja na jitihada zote hizo za uo
             Majeruhi hao, mmoja ni askari wa jeshi aliyekuw
la hiyo aliyetajwa kwa jina moja la Said naye amekimbizwa ho
kae chonjo, kwani mwanaume mmoja wa hapa Jijini aliyejifanya
kio lingine mtoto wa mwaka mmoja na nusu wa Ubungo-Kibangu,
                      Akalitaja gari lililohusika na ajali
                      Mtu mmoja Jijini ambaye anadaiwa kuwa
 hao pia wamemkata kidole kimoja na kuingia nacho mitini.
 ya usiku na askari polisi mmoja akiwa taabani huku kidole c
 Katika tukio lingine mkazi moja wa Sinza, Jiwa Shabani, 27,
             Ikiwa ni siku moja baada ya kusherehekewa kwa
staafu Ali Hassan Mwinyi na Meja Jenerali Herman Lupogo kuha
o wa kila siku, na sio siku moja kwa mwaka.
  Jumla ya Shilingi Milioni moja zimetumika kwa ajili ya kuw
 la KIWOHEDE Bi. Stela Mwambenja amesema wanawake walionufai
                      Akazitaja baadhi ya taka ngumu zinazo
funzo wanayoyatoa, Bi. Mwambenja akasema wanawake hao wamepa
```

```
i gheto huko Kariakoo na anakuja nyumbani kwa ajili ya kuiba
mkuu, viongozi, wachezaji pamoja na wanachama wa klabu" alis
irishwa kwa mchezo huo siku moja kabla bila sababu maalum.
umaa akasema tukio hilo limekuja baada ya kubainika kuwa yul
                  Mwanaume mmoja anayekisiwa kuwa na umri wa
 huohuo, maiti ya mwanaume mmoja imekutwa kandokando ya bara
Buguruni kuwa maiti ya mtu mmoja mwanaume imekutwa mapema le
o la kwanza, Afisa huyo alimtaja mtoto Prosper Simba toka ku
       Wakati huo huo, mtu mmoja Mhina Shabani mkazi wa kule
zi huo umefanyika katika viwanja vya Mnazi Mmoja na kuhudhur
a katika viwanja vya Mnazi Mmoja na kuhudhuriwa na wakazi we
                 Wanahusika moja kwa moja katika uchafuzi wa
a moja  Wanahusika moja kwa moja katika uchafuzi wa mazingir
ananchi kutunza taka sehemu moja na siyo kutupa ovyo.
wa kwenye kituo cha afya ni moja ya kero zilizotolewa kwake
madarakani kulikuwa na gari moja aina ya Landrover lakini ha
                       Mtu mmoja Jijini ambaye anadaiwa kuwa
 hao pia wamemkata kidole kimoja na kuingia nacho mitini.
```

The result has only some true hits. We construct constraints on the basis of the result (21).

(21)

```
$ cat-allmat | kwic 'ja ' | egrep -v '(o|n|i|e|a| |gu|[Mm]|[Nn])ja ' | wc
 10216
```

With the constraints in (21) we get quite a clean result, but the text is large for in-flight analysis. Here we have a case, where we must consider whether we are satisfied with the direct search plus constraints, or whether we go through both search phases, especially when the result of the first phase is already quite clean.
However, I have here implemented also the second phase. Part of the result is in (22).

(22)

```
 Anakuja {ja_V} Bush, Rais wa Marekani.
       -----
 Na ndio_maana awali tulisema kwamba, mawaidha yaweza kuwa mazuri, lakini
yakawa_na dosari kubwa, inapokuja {ja_V} nani kasema.
       -----
 Uzoefu unaonyesha kwamba, inapokuja {ja_V} jambo la Waislamu, mitulinga
imekuwa ndio mingi kwa_upande_wa serikali.
       -----
 Hapo_ndipo inapokuja {ja_V} ugumu wa kuzielewa nasaha za serikali kwamba
watu watumie hoja.
       -----
 Hakuja {ja_V} kwa_ajili_ya Waislamu bali kwa watu wote kwani hao ambao
si Waislamu wanaweza wakafaidika kwa kufuata mafunzo yake.
       -----
 Walikuja {ja_V} juu kuihami imani yao wasilishwe najisi ya nguruwe.
       -----
 Ali Mohamed Sheni alisema kuwa, kongamano hilo lilikuwa limekuja {ja_V}
katika muda muafaka na kwamba serikali itatekeleza yale yote yenye
umuhimu miongoni_mwa mapendekezo yatakayotolewa na wajumbe ikiwa ni njia
ya kutatua migogoro hiyo.
       -----
```

```
 Umuhimu wake unakuja {ja_V} kutokana_na ukweli kwamba sensa inasaidia
kupata taarifa muhimu katika kupanga maendeleo.
       -----
 Ni tarehe iliyoamsha hasira na chuki iliyokuja {ja_V} kutoa matunda
miaka ya mwanzo ya'.
       -----
 Vincenzo katika chapisho hilo alikuja {ja_V} na maelezo ya kutukuza
zaidi utaifa akishutumu wanaharakati wa nchi nyingine.
       -----
 Kwa msingi huo, lilipokuja {ja_V} suala la sensa, waliweka msimamo wao
kwa_uwazi kabisa.
       -----
 Ndio_maana kwasasa walikuja {ja_V} na hili la dhulma ya kukamatwa
viongozi wao na kuwekwa ndani kwa dhulma.
       -----
 Maoni na maelezo ya waislamu hao, yamekuja {ja_V} kufuatia kauli ya
Kaimu Mufti wa Bakwata, Sheikh Issa Shaaban, Mkuu wa Mkoa wa
Dar_es_Salaam na ile ya masheikh Kilemile, Juma Mikidadi na sheikh
Gorogosi.
       -----
 Yako mengi yanayoendelea kwa_siri, lakini nayo dhahiri yake itakuja
{ja_V} kujulikana siku moja.
       -----
 Yaweza kuwashangaza wengi, ni kwanini serikali imekuwa ikikiuka katiba
na sheria za nchi wakati inapokuja {ja_V} suala linalo wahusu waislamu.
       -----
 Serikali huja {ja_V} juu na kujipa mamlaka ya kusimamia na kudhibiti
kila linalowahusu waislamu na hata ikiwa ni uongozi basi uwe ule
unaokubalika na serikali.
       -----
 Hata_hivyo, haikuweza kufahamika ni nyenzo zipi zitakazotumika katika
uchaguzi huo wa kumpata kiongozi anayedaiwa atakuja {ja_V} kuwaongoza
waislamu.
       -----
 Yvonne anasema, ambacho amekuja {ja_V} kugundua ni kuwa, Uislamu ni dini
iliyo potoshwa kuliko nyingine yoyote.
       -----
 Wakati uamuzi wa kutaka kuwa Muislamu ni wake mwenyewe, anasema
imemshangaza sana jinsi watu wengine walivyokuja {ja_V} juu kutaka
kumzuia kuchukua uamuzi huo.
       -----
 Kusuasua huko kuna kuja {ja_V} kufuatia mahakama hiyo kushindwa
kusikiliza na kuiahirisha kutokana_na sababu mbali_mbali.
       -----
 Lakini, Wamarekani hawa wana anza kuja {ja_V} juu licha_ya kujua kwamba,
sheria hili linawashughulikia waislamu tena wale wakuja {ja_V}.
```

3 Conclusion

The experiments described in the report show that it is possible to do accurate search also from large text corpora, if we use a two-phase method. First, we do the normal string search using such a key, which covers the wanted words. The result, which contains mostly unwanted words, but also the wanted words, is then constrained, so that the result contains mostly wanted words. This greatly reduced text is then analysed and converted into the rich text format. In the second phase, the search is done from the rich text.

Because the rich text has a lemma form of each word, it is possible to get a clean result, which has only the wanted hits. The tests show that a 25 million word corpus is not too big for this kind of search, although the searched words are common in language. The most common words such as the verb *ku-wa* (to be) are too frequent for search even with the method described here.

In this report I have restricted to testing only non-extended verb forms of mono-syllabic verbs. I have also excluded such forms that end in *e* (subjunctive) or *i* (present tense negative).